ARMY RESEARCH LABORATORY

Fragmentation Modeling Using the Expectation Maximization Algorithm

by Andrew A. Thompson

ARL-TR-5508 **April 2011**

Approved for public release; distribution is unlimited.

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Aberdeen Proving Ground, MD 21005-5066

ARL-TR-5508 **April 2011**

Fragmentation Modeling Using the Expectation Maximization Algorithm

Andrew A. Thompson
Weapons and Materials Research Directorate, ARL

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED (From - To)
April 2011	Final	January 2010 - August 2010

4. TITLE AND SUBTITLE
Fragmentation Modeling Using the Expectation Maximization Algorithm

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Andrew A. Thompson

5d. PROJECT NUMBER
AH80

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-WML-A
Aberdeen Proving Ground, MD 21005-5066

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-5508

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The modeling of fractures typically uses the Poisson distribution to select the number of particles resulting from a fracture of a solid body. This applies to bullets when striking a plate. This report develops a method to model the number of subparticles using a mixture of Poisson distributions. The goal is to find the mixture with the smallest number of components that fits the data. Rather than attempting to find the distribution of λ, it is assumed the values of λ are clustered and the centroids of these clusters constitute a useful model.

15. SUBJECT TERMS
expectation maximization, Poisson mixtures, fragmentation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Andrew A. Thompson
Unclassified	Unclassified	Unclassified	UU	22	19b. TELEPHONE NUMBER (*Include area code*) 410-278-6805

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

Acknowledgments

The author would like to thank Dave Webb for the comments and suggestions he made that improved the quality of this report.

INTENTIONALLY LEFT BLANK.

1. Introduction

The modeling of fractures typically uses the Poisson distribution to select the number of particles resulting from a fracture of a solid body. This applies to bullets when striking a plate. If all the bullets are made by the same manufacturer using the same materials, then each bullet should have similar patterns of fragmentation. This statement is also true for similar manufacturers using similar material. In these cases, it is reasonable to expect a good fit of the fragmentation data to a single Poisson distribution. The single parameter λ of the distribution captures the fragmentation of the interaction. In practice and after an estimate is made, the fit to the data is checked both visually and through tests of fit.

For the situation where a bullet strikes a plate and then travels and strikes a second object, the use of a single Poisson distribution does not always fit the data. In this situation, each particle emerges from the first event, having undergone a unique experience of stress. In this situation, each subparticle should have its own λ to represent its rate of secondary fragmentation. This report develops a method to model the number of subparticles using a mixture of Poisson distributions. The goal is to find the mixture with the smallest number of components that fits the data. Rather than attempting to find the distribution of λ, it is assumed the values of λ are clustered and the centroids of these clusters constitute a useful model.

2. Background

The Expectation Maximization (EM) Algorithm was introduced by Dempster et al. in a seminal paper.[1] This paper provided a theoretical framework for many existing ad hoc approaches and fostered the development of many new estimation procedures that can be thought of in terms of missing data. The EM algorithm consists of two major steps. In the first step, the missing data is estimated based on the available information. This information is typically a combination of model parameters and observations. Expectation is used to estimate the unavailable data and form a complete set of data. The completed data is then used to find the maximum likelihood estimates of the model parameters. These two steps are repeated until an exit criterion is met. Part of the appeal of the EM algorithm is in the simplification of the maximization of the likelihood function associated with the specific problem; this simplification typically results from the formation of the complete data set. By approaching a problem from a more theoretical perspective, it is possible to find a solution; in many instances, the iterative algorithm is intuitive.

[1] Dempster, A. P.; Laird, N. M.; Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society B* **1977**, *39*, 1–38.

The estimation of mixtures has a long history extending back to Karl Pearson's paper[2] in which he estimated the mixture components of crab measurements as the mixture of two univariate normal distributions. His moments-based method was difficult to follow and required finding the proper root of a ninth-degree polynomial. While he was successful, many of his colleagues stated that only an individual of his intellectual stature was likely to be able to successfully apply his method. Although research continued in this area, it was not until the advent of the EM algorithm that mixtures of varied distributions were approachable.

In mixture modeling, parameter identification can be difficult. This difficulty arises because different mixtures of distribution functions can result in similar final distributions. For seemingly reasonable values of N, it may be difficult to distinguish between candidate mixtures. Often, this is referred to as an over parameterization problem. The ideas associated with observability of parameters based on the data can be used to discuss over parameterization; likewise, ideas associated with identifiability can also be applied to gain understanding. From the perspective of modeling, being ignorant of these possibilities can lead to zealousness associated with a particular mixture that appears to fit the data. Domain knowledge associated with the data can point out reasonable assumptions to make about the mixture and its components.

There have been many articles that investigate mixtures without appealing to the EM algorithm. These articles typically deal with a specific application and an interesting approach to the likelihood function. The EM algorithm allows a general approach that can be applied to a multitude of situations; thus, it is a useful technique for the prudent investigator. An excellent source of information on the EM algorithm was prepared by McLachlan and Krishnan.[3]

3. Algorithm Background

The distribution of a Poisson distribution contains one parameter, λ. This distribution is used to represent the number of occurrences of some event per unit. For example, the number of failures per year, the number of defects per yard, or the number of fragments per impact are all examples of the number of events per unit. Often, the number of times something happens per time unit is represented. For the current situation, the number of fractures per impact is represented. The distribution function for a Poisson is

$$f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}. \tag{1}$$

[2] Pearson, K. Contributions to the Mathematical Theory of Evolution. *Phil. Transactions of the Royal Society of London* A **1894**, *185*.

[3] McLachlan, G. F.; Krishnan, T. The EM Algorithm and Extensions; 2nd ed.; Wiley & Sons: New York, 2008.

The maximum likelihood estimate of λ is the sample mean of the data.

Consider a mixture of Poisson distributions; in this situation, the distribution would be

$$f(x,\Psi)=\sum_{i=1}^{g}\pi_i f(x;\lambda_i) \text{ where } \sum_{i=1}^{g}\pi_i=1, \tag{2}$$

where Ψ represents the parameters associated with the mixture concentrations and each distribution parameter. For the current situation, the parameters are the mixture proportions and the Poisson parameters. The likelihood function for this would be

$$L(\Psi|X)=\prod_{j=1}^{N}\left(\sum_{i=1}^{g}\pi_i f(x_j;\lambda_i)\right). \tag{3}$$

The difficulty with solving this likelihood function is associated with the summation. The product of N factors each containing g terms is daunting. Taking the log converts the product to a sum, but the log of a summation defies simplification. It would be a straightforward process to solve this if it was known which component distribution generated each data point. Unfortunately, this information is not available. If it were available, the data would be partitioned into g sets and the mean of each set would be the estimate for the λ of each component distribution. Complete data for this problem would include a variable indicating group membership. Let z_{ji} be an indicator variable for the jth observation belonging to the ith group. A value of 1 indicates membership. The likelihood function would then be

$$L(\Psi|X)=\prod_{j=1}^{N}\left(\sum_{i=1}^{g}\pi_i z_{ji} f(x_j;\lambda_i)\right). \tag{4}$$

Given complete knowledge, there would be a single one in each row of the matrix Z, and λ_i could be calculated as the mean of the ith group. This simple procedure to maximize the likelihood function is attractive; but z is not observable and, hence, is unknown. Since z_{ji} is unavailable, the expectation of z_{ji} is investigated.

4. Algorithm Development

As a first problem, consider estimating the mixture components π_i for the distribution $f(x,\Psi)$, where the individual Poisson parameters λ_i are known. The unknown parameters of the log likelihood function are the mixture components, π_i; thus, the partials with respect to the mixture

components are needed. All these components are positive and their sum is 1. They form a discrete probability distribution. For this case, the log likelihood of equation 3 is

$$log\left(L(\Psi|X)\right)=\sum_{j=1}^{N} log\left(\sum_{i=1}^{g}\pi_i f(x_j;\lambda_i)\right). \tag{5}$$

Since $1=\sum_{i=1}^{g}\pi_i$, it is possible to eliminate one of the parameters in equation 5; without loss of generality, let $\pi_g = 1-\sum_{i=1}^{g-1}\pi_i$, and then taking the partial of equation 5 with respect to a specific π_i leads to the following set of equations for $(i=1,\ldots,g-1)$ to be solved:

$$\sum_{j=1}^{N}\left\{\frac{f(x_j,\lambda_i)}{f(x_j,\Psi)}-\frac{f(x_j,\lambda_g)}{f(x_j,\Psi)}\right\}=0. \tag{6}$$

Since the denominator is the same for each term, it can be multiplied out of the equation. After reexamining equation 2, the reader may wish to convince their self that an explicit solution does not exist for the parameters, π_i. The direct application of maximum likelihood does not lead to a solution.

If the previously discussed indicator variable z_{ji} were known for this problem, the parameters could be computed as

$$\pi_i=\frac{\sum_{j=1}^{N} z_{ji}}{N}. \tag{7}$$

Knowledge of z_{ji} allows a simplification of the likelihood function. Consider the log of equation 4. The result seems similar to equation 5; however, the values of z_{ji} allow the log of the summation to be expressed as the log of a product.

$$log\left(L(\Psi|X,Z)\right)=\sum_{j=1}^{N} log\left(\sum_{i=1}^{g}\pi_i z_{ji} f(x_j;\lambda_i)\right). \tag{8}$$

The effect of the z_{ji} is to partition the inner sum based on mixture subcomponent, i.e., all subcomponents of $f(x,\Psi)$ not related to the ith index will be zeroed out of the expression. The variable Z allows the restructuring of the likelihood equation into a form that allows a closed form solution. After changing the order of the summations, the likelihood function can be expressed in an amenable form.

$$\begin{aligned} log\left(L(\Psi|X,Z)\right) &= \sum_{i=1}^{g} log\left(\sum_{j=1}^{N}\left(z_{ji}\pi_i f(x_j;\lambda_i)\right)\right) = \sum_{i=1}^{g}\left(\sum_{j=1}^{N} z_{ji}\, log\left(\pi_i f(x_j;\lambda_i)\right)\right) \\ &= \sum_{i=1}^{g}\left(\sum_{j=1}^{N} z_{ji}\, log\left(\pi_i\right)\right) + \sum_{i=1}^{g}\left(\sum_{j=1}^{N} z_{ji}\, log\left(f(x_j;\lambda_i)\right)\right). \end{aligned} \tag{9}$$

The right-most term of equation 9 will go to 0 when taking the partials with respect to the mixture parameters. This separation of the mixing parameters and the distribution parameters will allow the mixing parameters to be estimated without regard to the distribution parameters when the problem is made more general, as the existence of Z segments the mixing parameters and distribution parameters within the likelihood function.

It is also worth noting that the distributions do not need to be of the same family, for example, one component could be from a normal distribution and another from a Poisson. With this ability to mix different distributions comes the responsibility to make the choices of distribution based on domain knowledge. Domain knowledge will minimize the identifiability/observability problem by assisting the investigator in their choice of appropriate distributions.

Equation 7 allows the maximum likelihood solution of the mixing parameters. Since Z is not observable, its value will be based on expectation. For this problem, there is a reasonable approach to find the expectation of Z. The expectation of Z is found based on each observation. The expectation of z_{ji} is the probability that $\pi_i f(x_j, \lambda_i)$ was the component that generated the observation. The density value of an observation for each of the mixture component distributions can be found. Each of these values is the mixture value multiplied by the probability density value. The expectation of z_{ji} is described by the following formula:

$$E(z_{ji}) = \frac{\pi_i f(x_j;\lambda_i)}{f(x_j;\Psi)} \quad . \tag{10}$$

The denominator can be considered a normalization factor and ensures that the sum of the expectations associated with each observation is 1. Also note that the sum of all the z_{ji} will be N. After the values of z_{ji} are found, the estimated mixture proportions can be found via maximum likelihood as

$$\hat{\pi}_i = \frac{\sum_{j=1}^{N} z_{ji}}{N} \quad . \tag{11}$$

Calculations for the expectation and maximization are alternated until a convergence criterion is met. The final iterative algorithm is elegant in its simplicity and scope, and the contrast to Pearson's original investigation[1] is staggering.

The previously described algorithm will find the mixture proportions when the distribution parameters are known. For the situation of interest, the mixture proportions are unknown and must be estimated. It turns out that the introduction of the unknown membership variable, Z, also makes this a stand-alone step to add to the previous algorithm. Only the right-most term of equation 9 contains the distribution parameters.

$$\sum_{i=1}^{g}\left(\sum_{j=1}^{N} z_{ji}\, log\left(f(x_j;\lambda_i)\right)\right). \tag{12}$$

The maximum likelihood estimation of each Poisson parameter, λ_i, follows the maximum likelihood procedure and is demonstrated in many textbooks. Each λ_i can be found as the mean of its group.

$$\hat{\lambda}_i = \frac{\sum_{j=1}^{N} z_{ji} x_j}{\sum_{j=1}^{N} z_{ji}}. \tag{13}$$

The desired algorithm is formed by including the g computations associated with equation 13 to the maximization step of the previous algorithm. As previously mentioned, the selection of the number of components can be considered a subjective decision. In practice, it is possible to rerun the EM algorithm with an additional component and then stop when the percentage of an additional component is considered small or if two components have parameters that are almost equal. Initial component percentages can be based on prior information, data observation, or the minimum entropy (equally likely) assumption. The initial Poisson parameters can be selected. A valuable reference for application of the EM algorithm to mixtures has been prepared by McLachlan and Peel.[4]

5. Application of the Algorithm

The algorithm just described was instantiated in MATLAB code as a step or iteration. This routine was placed within a control structure that reran it until the values of the parameters were stable to four decimal places. Models with two to four components were run. The three-component model was chosen. The two-component model did not fit the data well; there was little difference between the outputs of the three- and four-component models. The four-component model had a component with a probability of less than 0.05. The strongest component of the selected model had a Poisson parameter of 0.1565 and a concentration of

[4] McLachlan, G. F.; Peel, D. *Finite Mixture Models*; Wiley & Sons: New York, 2000.

0.4691. Approximately 47% of the time the data had low probability of fracture. The second component had a probability of occurrence of 0.3557 and an average of 4.998 fractures. The third component occurred with a probability of 0.1752 and average 14.4648 fractures. For a Poisson distribution, the variability increased as the mean increased, so the third component had a large effect on the variability, even though it had the lowest probability of occurrence.

The actual data is displayed in figure 1. Figures 2–4 are realizations of the model. Visually, the model agrees with actual data. A Kolmogorov-Smirnov goodness-of-fit hypothesis test was performed to obtain a quantitative measure of the agreement between the actual data and a data set generated from the previously described mixture of Poisson distributions; the test failed to reject the hypothesis that the two sets came from the same distribution.

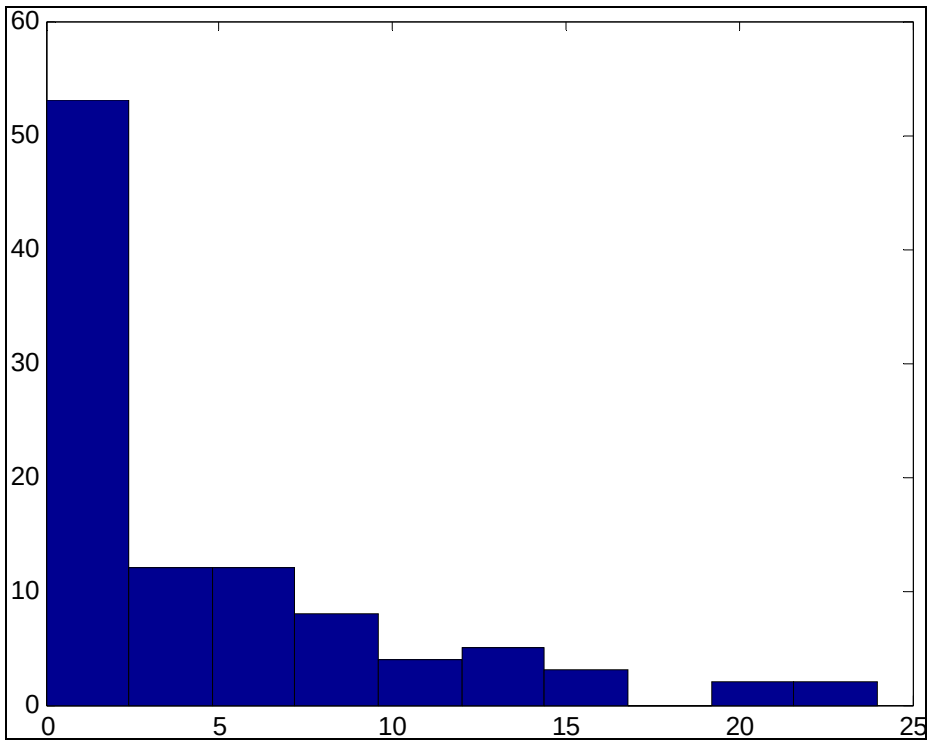

Figure 1. Actual data (number of fragments [x] vs. number of occurrences [y]).

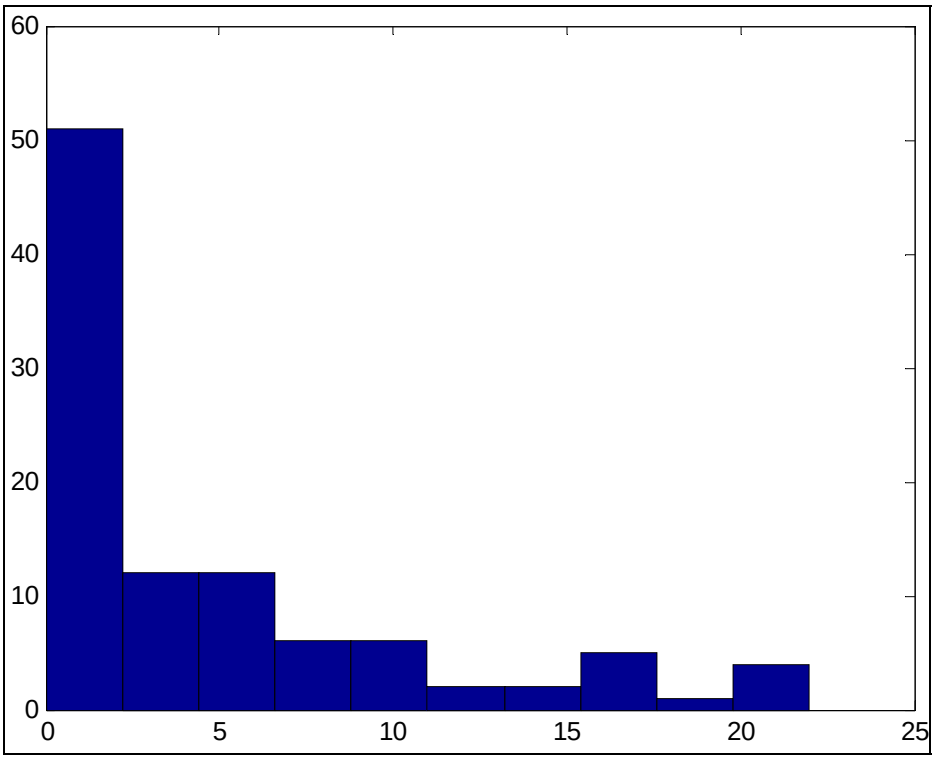

Figure 2. Realization 2 (number of fragments [x] vs. number of occurrences [y]).

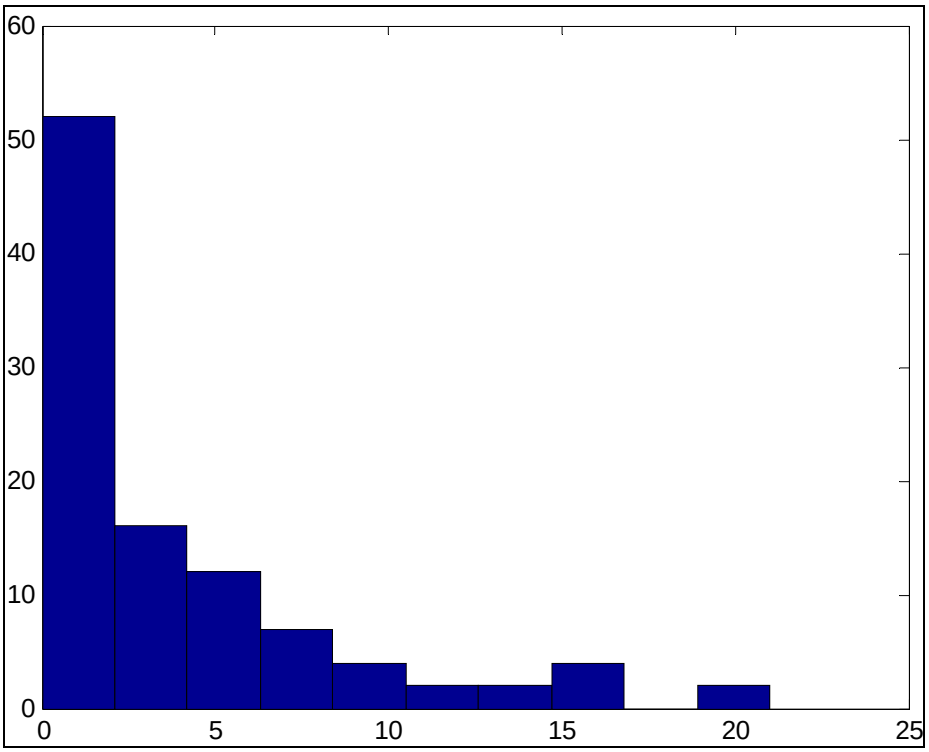

Figure 3. Realization of the mixture model (number of fragments [x] vs. number of occurrences [y]).

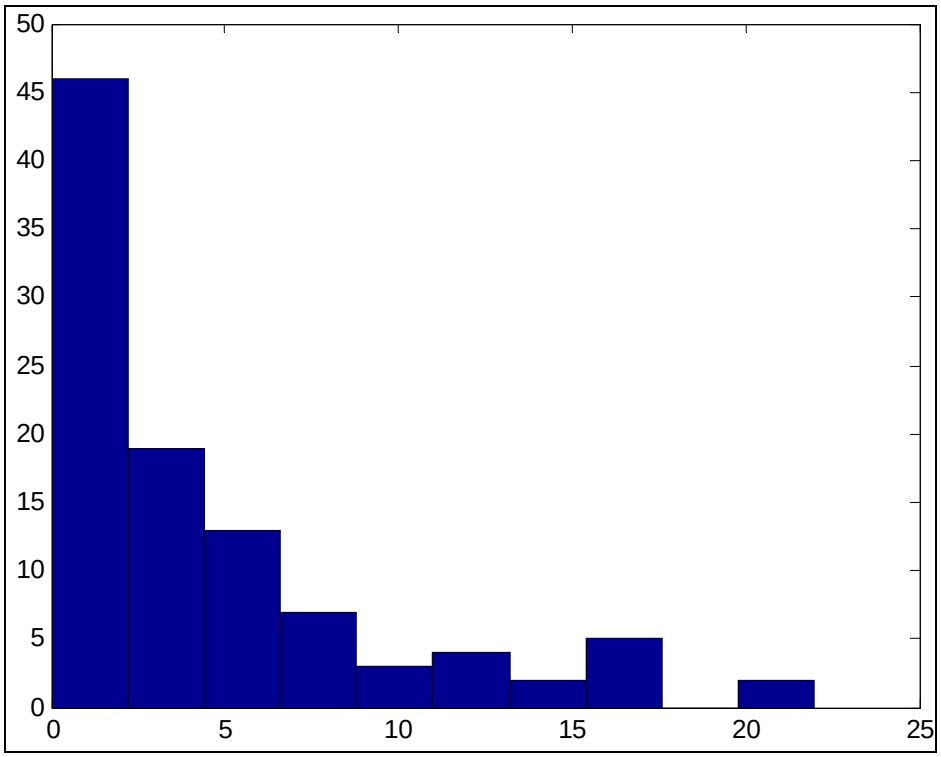

Figure 4. Realization 3 of the mixture model (number of fragments [x] vs. number of occurrences [y]).

6. Mixture of Poisson Regression Models

Generalized linear models (GLM) can be used within the framework of the expectation maximization algorithm to estimate the parameters for a mixture of Poisson regression models. Poisson regression refers to a class of models where the Poisson distribution parameter, λ, is estimated from a function of explanatory variables. Within the GLM framework, maximum likelihood is typically used to establish the relationship between the probability distribution and the explanatory variables. Using GLM within the EM algorithm requires the use of weights. These weights are analogous to those described by equation 10. For each observation, the normalized likelihood based on each GLM model is used as weight. The sum of the weights associated with a particular model divided by the number of observations gives the estimated mixture parameter. Mixtures of Poisson regression models are analogous to simple Poisson mixtures except that the parameter estimation step is more complicated. Estimation procedures for sophisticated models can be devised in a straightforward manner using the same general steps as those for a mixture of Poisson distributions.

7. Conclusions

Poisson models have been used to model fragmentation. In this report, the implementation of mixtures of Poisson models is discussed. These models provide valuable models for secondary fragmentation. In primary fragmentation where a bullet first strikes an object, the physical properties of the bullets are similar; however, after striking a plate, a bullet's physical properties can be altered and have large variability between bullets or fragments. The example used in this report demonstrated that it is possible to model this situation using a mixture of Poisson distributions. This mixture can be thought of as clusters of bullets or fragments with the same physical properties. The Kolmogorov-Smirnov test can be used to quantify the similarity between the data and the model.

If the information is available on predictor variables, Poisson regression can be used to model the fragmentation event. Models using striking velocity and angle information can be developed to increase modeling accuracy. Secondary fragmentation can be modeled using mixtures of Poisson regression models.

The investigator must always be aware of identifiability issues when utilizing mixture models. While a mixture model can be created to fit almost any data set, one should be skeptical of those with components not based on appropriate theory. For fragmentation data, Poisson or Poisson regression mixture models provide an approach that can increase the fidelity of the modeling effort.

NO. OF
COPIES ORGANIZATION

1 (PDF only) DEFENSE TECHNICAL INFORMATION CTR
DTIC OCA
8725 JOHN J KINGMAN RD
STE 0944
FORT BELVOIR VA 22060-6218

1 DIRECTOR
US ARMY RESEARCH LAB
IMNE ALC HRR
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM L
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL CIM P
2800 POWDER MILL RD
ADELPHI MD 20783-1197

1 DIRECTOR
US ARMY RESEARCH LAB
RDRL D
2800 POWDER MILL RD
ADELPHI MD 20783-1197

ABERDEEN PROVING GROUND

1 DIR USARL
RDRL CIM G (BLDG 4600)

NO. OF COPIES ORGANIZATION

2 GEORGIA TECH RSRCH INST
GTRI/ATAS
A LOVAS
SMYRNA GA 30080

1 DEPT MECHL ENGRG
DREXEL UNIV
B C CHANG
3141 CHESTNUT ST
PHILADELPHIA PA 19104

3 DEPT MECHL ENGRG
UNIV OF MARYLAND
A MOSLEH
M MODARRES
J HERRMANN
0151 F GLEN L MARTIN HALL
BLDG 088
COLLEGE PARK MD 20742

2 DARPA/I20
J FRITZGERALD
P MICHELUCCI
3701 N FAIRFAX DR
ARLINGTON VA 22203

2 DIRECTOR USARL
RDRL CI
B FORNOFF
D DENT
2800 POWER MILL RD
ADELPHI MD 20783

1 DIRECTOR USARL
RDRL CII
B BROOME
2800 POWER MILL RD
ADELPHI MD 20783

1 DIRECTOR USARL
RDRL CII A
C PIERCE
2800 POWER MILL RD
ADELPHI MD 20783

1 DIRECTOR USARL
RDRL CII B
L TOKARCIK
2800 POWER MILL RD
ADELPHI MD 20783

NO. OF COPIES ORGANIZATION

1 DIRECTOR USARL
RDRL CIH N
C ADAMS
2800 POWER MILL RD
ADELPHI MD 20783

1 DIRECTOR USARL
RDRL CIN D
A CLARK
2800 POWER MILL RD
ADELPHI MD 20783

1 DIRECTOR USARL
RDRL CIN S
C ARNOLD
2800 POWER MILL RD
ADELPHI MD 20783

2 DIRECTOR USARL
RDRL CIN T
B RIVERA
R HARDY
2800 POWER MILL RD
ADELPHI MD 20783

ABERDEEN PROVING GROUND

25 DIR USARL
RDRL CII C
B BODT
A NEIDERER
J DUMER
RDRL CIH
R NAMBURU
RDRL CIH M
D WILSON
RDRL CIN D
C ELLIS
L MARVEL
R RESCHLY
RDRL SLB D
J COLLINS
L MOSS
RDRL SLB W
P GILLICH
RDRL WML A
J WALL
D WEBB
M ARTHUR
A THOMPSON (4 CPS)
B FLANDERS
R PEARSON
B OBERLE

NO. OF
COPIES ORGANIZATION

RDRL WML F
M ILG
R MCGEE
T HARKINS
RDRL WML G
T BROWN

INTENTIONALLY LEFT BLANK.